Mining relations in the GENIA corpus

Rinaldi, Fabio ; Schneider, G ; Kaljurand, K ; Dowdall, J ; Andronis, C ; Persidis, A ; Konstanti, O

Abstract: Discovering the interactions between genes and proteins is seen as one of the core tasks in molecular biology. The quantity of research results in this area is growing at such arate that it is very dicult for individual researchers to keep track of them. As such results appear mainly in the form of scientic articles, it is necessary to process them in an ecient manner in order to be able to extract the relevant results. Many databases exist that aim at consolidating the newly gained knowledge in a format that is easily accessible and searchable, however the creators of such databases normally make use of human readers who manually curate the rel- evant papers. This is an expensive and time consuming process, besides, there might be a signicant time lag be- tween the publication of a result and its introduction into such databases. In this paper we propose a method for discovery of inter- actions between genes and proteins from the scientic liter- ature, based on a complete syntactic analysis of the corpus. We report on preliminary results.

Mining relations in the GENIA corpus

Fabio Rinaldi, Gerold Schneider,
Kaarel Kaljurand , James Dowdall*
Institute of Computational Linguistics,
University of Zurich,
CH-8057 Zürich, Switzerland
http://www.cl.unizh.ch/
{rinaldi,gschneid,kalju,dowdall}@ifi.unizh.ch

Christos Andronis, Andreas Persidis,
Ourania Konstanti
Biovista, 34 Rodopoleos Str,
Ellinikon, GR-16777 Athens
Greece
http://www.biovista.com/
{candronis, andreasp, okonst}@biovista.com

ABSTRACT

Discovering the interactions between genes and proteins is seen as one of the core tasks in molecular biology. The quantity of research results in this area is growing at such a rate that it is very difficult for individual researchers to keep track of them. As such results appear mainly in the form of scientific articles, it is necessary to process them in an efficient manner in order to be able to extract the relevant results.

Many databases exist that aim at consolidating the newly gained knowledge in a format that is easily accessible and searchable, however the creators of such databases normally make use of human readers who manually 'curate' the relevant papers. This is an expensive and time consuming process, besides, there might be a significant time lag between the publication of a result and its introduction into such databases.

In this paper we propose a method for discovery of interactions between genes and proteins from the scientific literature, based on a complete syntactic analysis of the corpus. We report on preliminary results.

1. INTRODUCTION

One of the core problems in exploiting scientific papers in research and clinical settings is that the knowledge that they contain is not easily accessible. Although various resources which attempt to consolidate such knowledge are being created (e.g. UMLS[1], SWISS-PROT, OMIM, GeneOntology, GenBank, LocusLink), the amount of information available keeps growing exponentially [29].

Besides, the creation of such resources is a very labour intensive process. Relevant articles have to be selected and accurately read by an human expert looking for the relevant information.[2]

The various genome sequencing efforts have resulted in the creation of large databases containing gene sequences. However such information is of little use without the knowledge of the function of each gene and its role in biological pathways. The study of the interactions within and between genes and proteins form a key part of research activities in the domain of molecular biology.

A gene contains hereditary information encoded in the form of DNA and is located at a specific position on a chromosome in a cell's nucleus. Genes determine many aspects of anatomy and physiology by controlling the production of proteins (gene products). Gene products form interconnected networks in order to accomplish specific goals. A biological process (*pathway*) is accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are "*cell growth and maintenance*" or "*signal transduction*". Examples of more specific terms are "*pyrimidine metabolism*" or "*alpha-glucoside transport*". Understanding the relationships within and between these groups is central to biology research and drug design as they form an array of intricate and interconnected molecular interaction networks which is the basis of normal development and the sustenance of health.

One of the problems in this task is that current understanding of biology exists in islands of knowledge which are often ill connected. In recognition of this situation a number of approaches are currently being developed in order to help with the generation of hypotheses which can later be confirmed or refuted in wet lab experiments. Literature-based Discovery (LBD) is one such approach that uses free text (scientific articles) as its raw material.

There are various commercial tools which aim at supporting the LBD process. One example is Biovista's BioLab Experiment Assistant (BEA),[3] which is a literature-based environment that supports researchers in exploring problem areas of their choice, discovering hidden links and designing their experimental strategy. By integrating and cross correlating a number of research parameters (such as genes, pathways, diseases and cell lines), BEA provides multidimensional coverage of the life sciences domain and supports users in hypothesis generation and experiment strategy design in a comprehensive manner.

We describe an approach to the extraction of such relations from domain corpora based on a full parsing of the documents and on a set of rules that map syntactic structures into the relevant relations. In section 2 we describe the

*Now at the Department of Informatics, University of Sussex, UK.

[1] http://www.nlm.nih.gov/research/umls/

[2] This process is referred to as 'curation' of the article.

[3] BEA is a trademark of Biovista. All rights reserved.

In *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics*, held in conjunction with ECML/PKDD in Pisa, Italy. 24 September 2004.

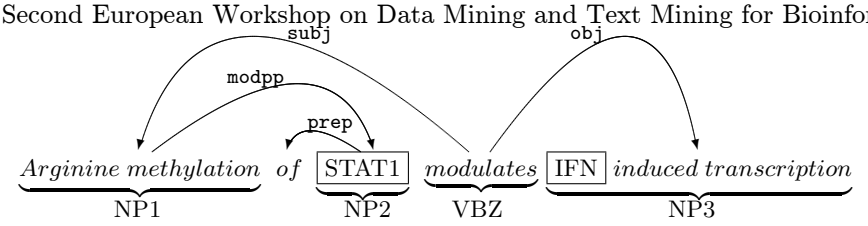


Figure 1: An example of syntactic analysis

nature of the corpus that we have adopted as a testbed for our work. In section 3 we describe our methodology, which we then evaluate in section 4. In section 5 we discuss some advanced applications and our plans for future activities, in section 6 we survey related work.

2. THE CORPUS

We are basing our experiments on two different collections in the domain of molecular biology. The first collection (here called the 'raw' corpus) has been generated from Medline using Biovista's BEA system, using two seed term lists of genes and pathways. The second collection is constituted by the GENIA corpus [11].[4]

BioLab Experiment Assistant (BEA) is a literature-search and analysis tool that supports researchers in exploring problem areas of their choice, discovering hidden links and formulating research hypotheses in the life sciences domain. The corpus of BEA is a combination of more than 50.000 full text articles as well as over 13 million abstracts from Medline. By extracting and cross correlating over 250.000 life sciences related terms organised in 11 taxonomies/ontologies, BEA creates a weighted co-occurrence network that is used to navigate, and analyse large literature collections in a systematic and intuitive manner.

In the case of the raw corpus, we had to perform a phase of terminology discovery, which was facilitated by the existence of the seed lists of genes and pathways [5]. We first marked up those terms which appear in the corpus using additional XML tags. This identified 900 genes and 218 pathways that occur in the corpus. (represented as boxed tokens in fig. 1). Next the entire corpus is chunked into nominal and verbal chunks using LT Chunk [6]. Ignoring prepositions and gerunds, the chunks are a minimal phrasal group (represented as the square braces in fig. 1). The corpus terms are then expanded to the boundary of the phrasal chunk they appear in. For example, NP3 in fig. 1 contains two terms of interest producing the new term "*IFN-induced transcription*". The 1118 corpus terms were expanded into 6697 new candidate terms. 1060 involve a pathway in head position and 1154 a gene. The remaining 4483 candidate terms involve a novel head with at least one gene or pathway as a modifier.

We have described in [17] some approaches that might be taken towards terminology extraction for a specific domain. The GENIA corpus removes these problems completely by providing pre-annotated terminological units. This allows attention to be focused on other challenges, rather than getting 'bogged down' with terminology extraction and organization. Although the problem of detecting domain-specific entities is a crucial one, as the focus of this paper is on detecting relations, we will mainly refer to the GENIA corpus. The 'raw' corpus is only been used in the application described in section 5.

We use version G3.02 of the GENIA corpus. There are 2000 articles[5], 18546 sentences (average length 9.27 sentences per article), 490941 words (average of 26.47 words per sentence). The GENIA corpus has been annotated for various biological entities, according to the GENIA Ontology.[6]

3. DESCRIPTION OF THE EXPERIMENTS

In a first step, we convert the XML annotations of the GENIA corpus into an equivalent annotation schema defined within the scope of the Parmenides project [15]. There are two main reasons for performing this step. First, in the Parmenides annotation schema all relevant entities are given a unique identifier. As identifiers are preserved during all steps of processing, the existence of a unique identifier for each sentence and each token in the corpus later simplifies the task of presenting the results to the user. The second reason is that the Parmenides annotation scheme allows for a neater distinction of different 'layers' of annotations (structural, textual and conceptual) which again simplifies later steps of processing.

In a second step, we chunk the GENIA corpus using LT Chunk, and we create a merged version of the corpus, containing both the original GENIA annotations and additional XML annotations to mark chunk boundaries (this is later on called the **'clean' corpus**).

As a first result, we want to show that the availability of domain terminology simplifies and improves the task of parsing the corpus. To this aim we create a **'dirty' corpus**, which contains the chunk boundaries detected at the previous step, but does not contain the original GENIA markup for domain terminology.

Second, we want to verify whether the parsing of the corpus can benefit from the existence of semantic tags. The idea is to allow the parser to decide on an ambiguous attachment based on the semantic type of the arguments. For instance the decision of attaching an argument of type 'protein' as the subject of the verb 'bind' could be made on the basis of the type, rather than based purely on the lexical item itself. This approach is inspired by [9] and [30], the former extracting unambiguous but sparse PP-attachment information from the unannotated GENIA corpus, the latter also using ambiguous but less sparse PP data.

The third result that we describe in this paper concerns the detection of specific relations by means of specific lexical classes and a small set of rules that describe specific syntactic patterns. This can be seen as partly similar to [13], which

[4] `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`

[5] Actually 1999, because article number 97218353 appears twice, curiously with slightly different annotations.

[6] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html`

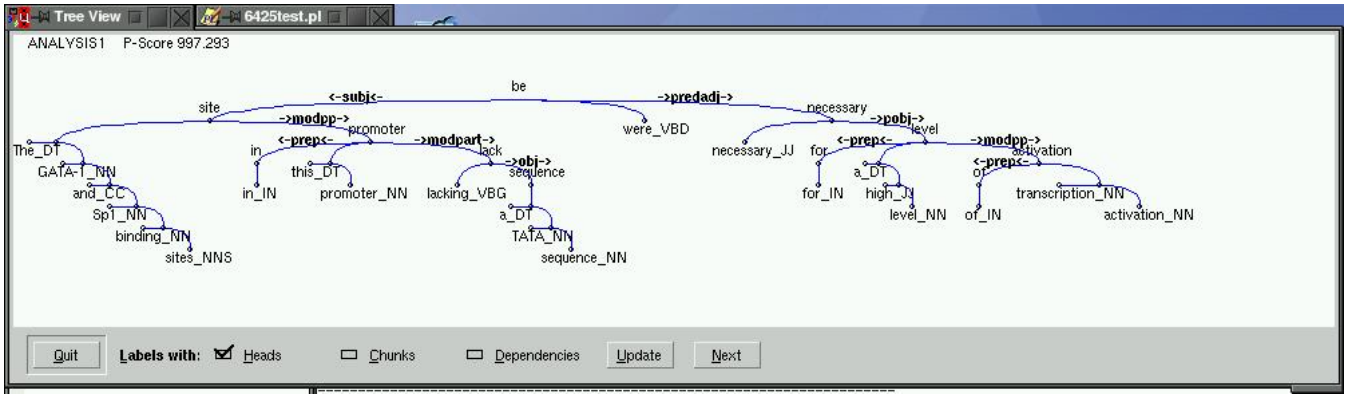


Figure 2: Dependendency Tree output of of the SWI Prolog graphical implementation of the parser

```
subj(bind, ngfi-b/nur77, _, '<-').
prep(element, to, _, '<-').
prep(monomer, by, _, '<-').
conj(heterodimer, or, _, '<-').
prep(receptor, with, _, '<-').
appos(receptor, rxr, _, '->').
modpp(heterodimer, receptor, with, '->').
conj(monomer, heterodimer, or, '->').
modpp(element, monomer, by, '->').
pobj(bind, element, to, '->').
```

Table 1: Parse of sentence 3800

however makes use of surface POS-based patterns, while our patterns apply to the result of syntactic parsing.

As an example consider GENIA sentence number **3800**:[7] *NGFI-B/nur77* **binds to** *the response element by monomer or heterodimer with retinoid X receptor (RXR).*

Based on the interaction with a domain expert, we have identified a set of relations that are of particular interest in this domain. Some examples of relevant relations are: *activate, bind, interact, regulate, encode, signal* [7]. We have then expanded the 'seed words' with their morphological variants (e.g. *bind* → *bind, binds, binding, bound*).

For each of those relations, we have inspected some of the analysis that we obtained from parsing the corpus, such as the one shown in table 1 for the example sentence 3800.

On the basis of such inspection, we developed a number of axioms that capture the relations that are of interest in this domain. For example, in the case of the *binds to* expression, the **bind** relation can be captured by the following axiom:

```
subj(bind,X,_,_),pobj(bind,Y,to,_)
prep(Y,to,_,_) => bind(X,Y).
```

Before presenting the results obtained using this method (section 4), we describe the core component of our approach: the Pro3Gres parser.

3.1 Parsing

The deep syntactic analysis builds upon the chunks using a broad-coverage probabilistic Dependency Parser [24] to identify sentence level syntactic relations between the heads of the chunks. The output is a hierarchical structure of syntactic relations - functional dependency structures, represented as the directed arrows in fig. 1. The parser (*Pro3Gres* [24, 26]) uses a hand-written grammar combined with a statistical language model that calculates lexicalized attachment probabilities, similar to [3]. Parsing is seen as a decision process, the probability of a total parse is the product of probabilities of the individual decisions at each ambiguous point in the derivation.

Two supervised models (based on Maximum Likelihood Estimations (MLE)) are used. The first is based on lexical probabilities of the heads of phrases, calculating the probability of finding specific syntactic relations (such as subject, sentential object, etc.). The second probability model is a Probabilistic Context Free Grammar (PCFG) for the production of verb phrases. Although Context Free Grammars (CFG) are alien to dependency grammar, verb phrase PCFG rules can model verb subcategorization frames which are an important component of a dependency grammar.

Probabilistic parsers generally have the advantage that they are fast and robust, and that they resolve syntactic ambiguities with high accuracy. Both of these points are prerequisites for a statistical analysis that is feasible over large amounts of text.

In comparison to shallow processing methods, parsing has the advantage that relations spanning long stretches of text can still be recognized, and that the context largely contributes to the disambiguation.

In comparison to deep linguistic, formal grammar-based parsers, however, the output of probabilistic parsers is relatively shallow, pure CFG constituency output, i.e. tree structures that do not express long distance dependencies (LDDs). In a simple example "*John wants to leave*" a shallow CFG analysis does not express the fact that John is also the implicit subject of *leave*. A parser that fails to recognize these implicit subjects, so-called control subjects, misses very important information, quantitatively about 3% of all subjects.

The parser expresses distinctions that are especially important for a predicate-argument based shallow semantic representation, as far as they are expressed in the Penn Treebank training data, such as PP-attachment, most LDDs, relative clause anaphora, participles, gerunds, and the argument/adjunct distinction for NPs.

In some cases functional relations distinctions that are

[7] In the original GENIA corpus sentences are not numbered, numbers are assigned during conversion to the Parmenides format.

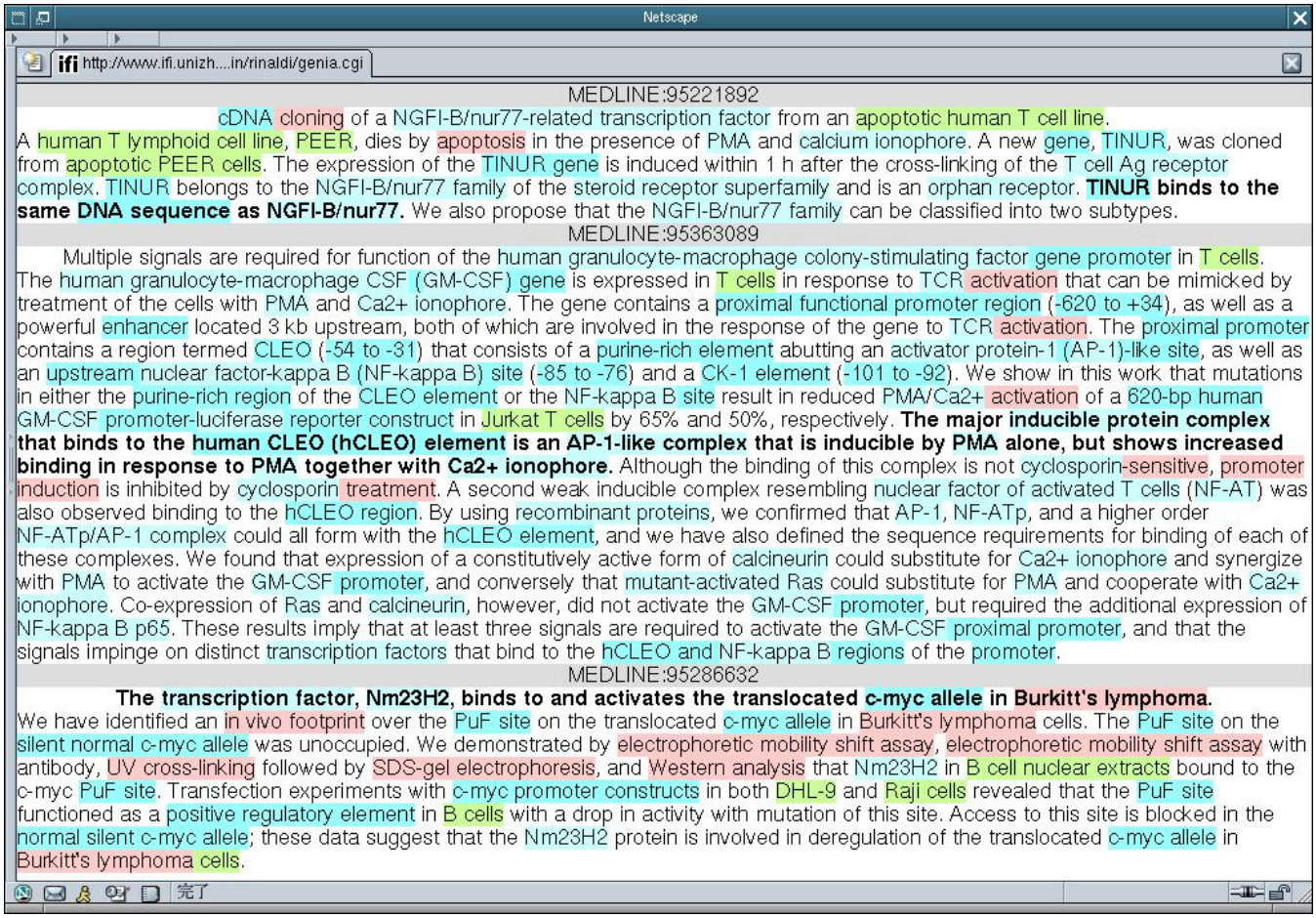


Figure 3: Inspecting relevant sentences

Relation	Label	Example
verb–subject	subj	*he sleeps*
verb–first object	obj	*sees it*
verb–second object	obj2	*gave (her) kisses*
verb–adjunct	adj	*ate yesterday*
verb–subord. clause	sentobj	*saw (they) came*
verb–prep. phrase	pobj	*slept in bed*
noun–prep. phrase	modpp	*draft of paper*
noun–participle	modpart	*report written*
verb–complementizer	compl	*to eat apples*
noun–preposition	prep	*to the house*

Table 2: The most important dependency types used by the parser

not expressed in the Penn Treebank are made. Commas are e.g. disambiguated between apposition and conjunction, or the Penn tag *IN* is disambiguated between preposition and subordinating conjunction. Other distinctions that are less relevant or not clearly expressed in the Treebank are left underspecified, such as the distinction between PP arguments and adjuncts, or a number of types of subordinate clauses.

The parser is robust in that it returns the most promising set of partial structures when it fails to find a complete parse for a sentence. A screenshot of its graphical interface can be seen in fig. 2. Its parsing speed is about 300,000 words per hour. Initial results of parsing the GENIA corpus are reported in [25]. More complex applications (Question Answering) are described in [20, 19].

4. EVALUATION

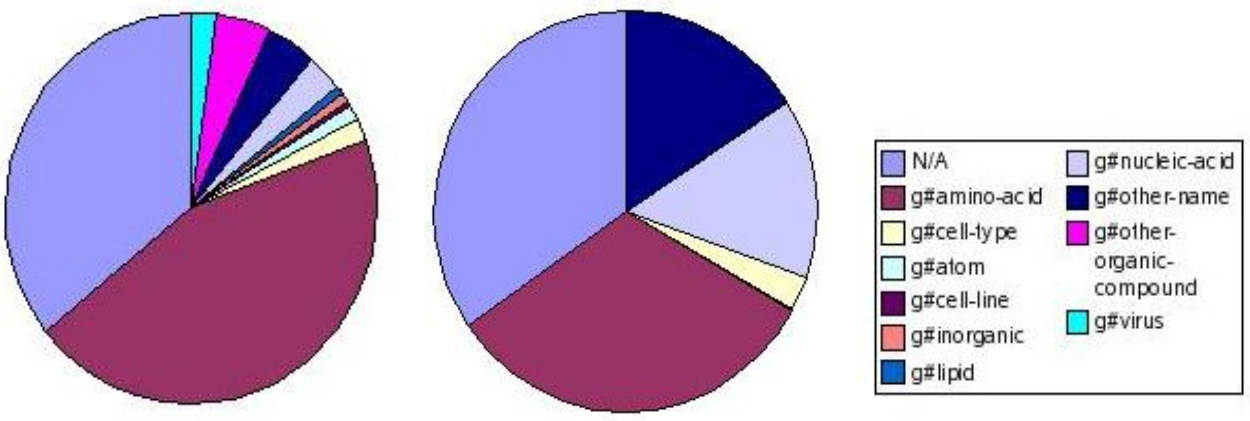
Two different types of evaluation have been performed. First a linguistic evaluation of the parser. Next we focused on the evaluation of the biological significance of the extracted relations.

4.1 Parser Evaluation

In order to perform an evaluation on the various experiments mentioned in the previous section we have randomly selected 100 sentences from the GENIA corpus, which we have manually annotated for the syntactic relations that the parser can detect.

We have first run the parser over the 100 sentences as extracted from the 'dirty' corpus, containing the chunks as generated by LTCHUNK, but no information on terminology. Later we have performed the analysis over the same 100 sentences, however this time extracted from the 'clean' corpus. A comparison of the results is shown in table 3.

More experimentally, we have integrated PP-attachment modules [9, 30] using the GENIA corpus, because the original PP-training corpus (the Penn Treebank) is of a different domain. Against sparse data we back off to semantic GENIA



a) Subject b) Object Legend

Figure 4: Distribution of arguments for the 'activate' relation

classes. Our current results do not show any improvement.[8]

Detailed results of the evaluation of the parser are reported in [25, 26].

4.2 Relation Evaluation

So far we have focused on triples of the form (predicate - subject - object).[9] The analysis of the whole GENIA corpus resulted in 10072 such triples (records). For the evaluation of biological relevance we selected only the records containing the following predicates: *activate, bind* and *block*. This resulted in 487 records.

The extraction algorithm aims at maximally expanding the arguments of the predicate, following all their dependencies. Each argument is then assigned a type (a concept of the GENIA Ontology), based on its head. The type assignment depends on the manual annotation performed by the GENIA annotators, so we have taken it as reliable and have not further evaluated it. We then removed all records where a type had not been assigned to either subject or object: this left 169 fully qualified records.[10] This remaining set was inspected by a domain expert.

In order to simplify the process of evaluation, we have created simple visualization tools (based on XML, CSS and CGI scripts), that can display the results in a browser. For instance, for the former type of evaluation, our visualization tool adds a special attribute to the sentences that have been detected by the methodology previously described. All the articles that contain relevant sentences are then automatically collected and displayed in a browser. We have slightly modified the original CSS provided with GENIA, so that only the relevant sentences are displayed in boldface (see fig. 3). The extracted relations can also be stored in a DB format for further processing with a spreadsheet tool or for analysis with Data Mining algorithms. For example, fig. 4 shows the distribution across the various GENIA types of

[8]This might be attributed to insufficient data or the relative simplicity of the GENIA Ontology.

[9]Which amounts to use axioms of the form: `subj(V,X,_,_),obj(V,Y,_,_) => V(X,Y)`

[10]This step is meant to remove records where one of the arguments cannot be clearly assigned a type. This is generally caused by pronouns, which explains why in the error evaluation (see table 5) the number of pronouns appears so low.

Relation	dirty	clean	semantic
subj (precision)	0.825	0.900	0.888
subj (recall)	0.744	0.862	0.846
obj (precision)	0.701	0.941	0.941
obj (recall)	0.772	0.949	0.949
sentobj (precision)	0.630	0.711	0.692
sentobj (recall)	0.604	0.75	0.729

Table 3: Comparison of results of parsing under different conditions.

Subjects and Objects for the `activate` relation.

The first 'naive' evaluation was based on assigning a simple key code to each record: 'P' for positive (biologically relevant and correct, 53 cases), 'Y' for acceptable (biologically relevant but not completely correct, 102 cases) and 'N' (not biologically relevant or seriously wrong, 14 cases). This result was considered as encouraging as it showed 91.7% of relevant records.

On closer inspection of the results reported by the domain expert, we identified a number of 'typical cases', which we then asked the expert to evaluate in detail. In this second evaluation the expert had to evaluate each argument separately and mark it according to the following codes:

- **Y** the argument is correct and informative
- **N** the argument is completely wrong
- **Pr** the argument is correct, but it is a pronoun, and it would need to be resolved to be significant (e.g. "This protein").
- **A+** the argument is "too large" (which implies that a prepositional phrase has been erroneously attached to it)
- **A-** the argument is "too small" (which implies that an attachment has been omitted)

In table 4 we show as an example the evaluation of the following sentences:

178. *Interleukin-2 (IL-2) rapidly activated Stat5 in fresh PBL, and Stat3 and Stat5 in preactivated PBL.*

807. *Thus, we demonstrated that IL-5 activated the Jak 2 -STAT 1 signaling pathway in eosinophils.*

5212. *Spi-B binds DNA sequences containing a core 5-GGAA-3 and activates transcription through this motif.*

No	relation	subj	subj type	subj eval	obj	obj type	obj eval
178	activate	Interleukin-2 (IL-2)	G#amino_acid	Y	Stat5 in fresh PBL, and Stat3 and Stat5 in pre-activated PBL	G#amino_acid	A+
807	activate	IL-5	G#amino_acid	Y	the Jak 2 -STAT 1 signaling pathway	G#other_name	Y
5212	bind	Spi-B	G#amino_acid	Y	DNA sequences	G#nucleic_acid	A-
16919	bind	The higher affinity sites	G#other_name	Pr	CVZ with 20-	G#other-_organic-_compound	N

Table 4: Some examples of evaluation.

	Y	N	Pr	A+	A-
Subject	146	11	4	6	2
Object	99	1	4	59	6

Table 5: Distribution of errors

16919. *The higher affinity sites bind CVZ with 20- to 50-fold greater affinity, consistent with CVZ's enhanced biological effects.*

We then noticed that some of the relations that had originally been considered as negative, had to be reconsidered, because our algorithm at present does not detect polarity (e.g. *"does not activate"*) or modality (e.g. *"might activate"*) and therefore some of the negative or hyphotetical cases, which the domain expert considered as incorrect, should be accepted for the purpose of the present evaluation.

Once all those points were clarified, we repeated the evaluation, which resulted in the values shown in table 5. This clearly shows that the biggest source of error is overexpansion of the object, plus there is a little but not insignificant problem in the detection of the subject.[11] Despite the errors, the results can be considered satisfactory, as they show 86.4% and 58.6% correct results in the detection of subjects and objects (respectively). If all loose cases are considered as positive (excluding only the 'N' cases), these results jump to 93.5% and 99.4% (respectively).

5. ADVANCED APPLICATIONS

The techniques described in this paper are currently being refined and expanded. We will first correct the parsing errors that have been identified by the present evaluation, then we will add facilities for the detection of the polarity and the modality of the relation. Another urgent task is the treatment of nominalizations (e.g. *"activation"*)[12] and other morphological transformations of the relations of interest (e.g. *"activators", "the activated protein , "co-activation"*), which are currently ignored. Further, some spelling variants should be considered (e.g. *"analyze"* vs. *"analyse"* or *"down-regulate"* vs. *"downregulate"*). We also want to expand the set of axioms in order to detect more complex relations.

A larger experiment that we are considering involves analysis of non manually annotated Medline documents. We will use the BEA tool to select a number of significant Medline articles, then we will use BEA's internal resources to automatically annotate them for relevant biological entities. BEA is an integrated knowledge environment that uses information extraction techniques combined with ontologies to organise biology and life sciences-related knowledge into weighted co-occurrence networks covering 50+ thousand full text articles and 13+ million abstracts in Medline. The heart of BEA is a database of concepts cross-correlated on the basis of their co-occurrence within the full text of scientific articles found in the top Science Citation Index biotech and medical journals. Currently BEA extracts and correlates the following concept classes: genes, pathways, post translational modifications, diseases, cell lines, organisms, experimental procedures, reagents, medical tests and authors. In addition to this, the BEA database contains all patents in the health-related categories. The work described in this paper is being considered as a possible expansion of the BEA system: the correlations discovered on the basis of statistical co-occurrence could be 'precised' on the basis of the linguistic analysis of the documents.

Another advanced application that we are considering is in a Question Answering system over scientific literature in the domain of Genomics [19]. ExtrAns is a QA system specifically targeted at technical domains [20], which can make intelligent use of available resources for a given domain, in particular terminology and ontology. The high frequency of terminology in technical text produces various problems for NLP applications. A primary problem is the increased difficulty of parsing text in a technical domain due to domain-specific sublanguage. Various types of multi-word terms characterize these domains, in particular referring to specific concepts (e.g. genes, proteins). Not only the internal structure of the compound can be multi-way ambiguous, also the boundaries of the compounds are difficult to detect and the parser may try odd combinations of the tokens belonging to the compounds with neighboring tokens.

It becomes crucial therefore to identify reliably all terminology of the domain. Once the terminology is available, it is necessary to detect relations among terms in order to exploit it. We have focused our attention in particular to the relations of synonymy and hyponymy, which are detected as described in [16] and gathered in a Thesaurus. More complex structuring (and even mapping to an existing domain Ontology) can be achieved using the techniques presented in [23]. The organizing unit of our Thesaurus is the WordNet style synset which includes strict synonymy as well as three weaker synonymy relations. These sets are further organized into a isa hierarchy based on two definitions of hyponymy.

The Question Answering system can then make intelligent use of such relations, and retrieve not only the answers that

[11] A close inspection of these cases points to problems with conjunctions in subject position, plus a specific problem with the construction *"does not"*.

[12] A simple inspection shows that "activation" makes up almost 50% of the occurrences of the stem "activat*".

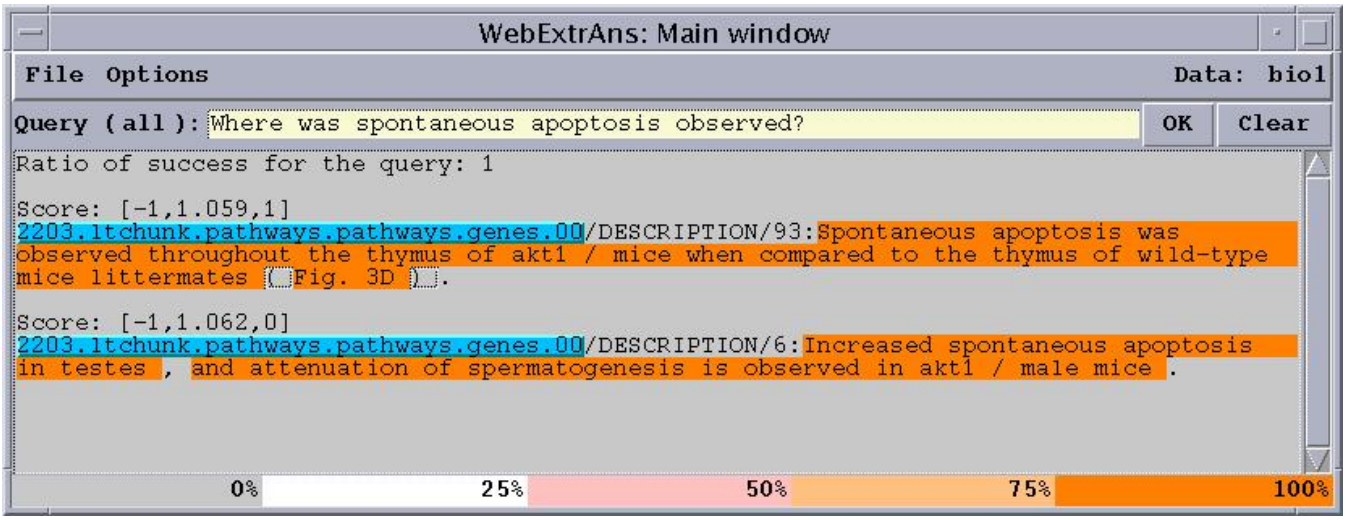


Figure 5: Example of usage of the QA system

contain the same term that was mentioned in the query, but also the related terms (paraphrases), as described in [18]. An example of Question Answering over the raw corpus can be seen in fig. 5.

6. RELATED WORK

While a majority of the applications of Natural Language Processing Techniques in the domain of molecular biology tend to focus on Entity Discovery, such as Genes and Proteins (see for instance [10] and [1]) there are some significant works in detecting relations among those entities.

For example, [4] identifies possible drug-interaction relations (predicates) between proteins and chemicals using a 'bag of words' approach applied to the sentence level. This produces inferences of the type: drug-interactions (protein, pharmacologic-agent) where an agent has been reported to interact with a protein.

[13] reports on extraction of protein-protein interactions based on a combination of syntactic patterns. The authors employ a simple dictionary lookup procedure to identify proteins in the documents to analyze, then select sentences that contain at least two proteins, which are then parsed with very simple part-of-speech matching rules. The rules are triggered by a set of (stemmed) keywords which are frequently used to name protein interactions (e.g. 'associate', 'bind', etc.) and can identify negative statements (again by matching specific words, such as 'not').

Methods partially similar to those that we adopted have been presented in [7]. They describe a system (GENIES) which extracts and structures information about cellular pathways from the biological literature. [14] processes Medline articles (only titles and abstracts) focusing on relation identification. An advantage of their system is the anaphora resolution module, which can resolve many cases of pronominal anaphora and anaphora of the sortal type (e.g. *"the protein"*) including multiple antecedents (e.g. *"both enzymes"*). Their evaluation is based on the `inhibit` relation.

The PASTA system [8] uses a template-based Information Extraction approach, focusing on the roles of specific amino acid residues in protein molecules. Similar to our approach is the usage of syntactic analysis resulting in a predicate argument representation. On the basis of such representation they also build a domain model which allows inferences based on multiple sentences.

[27] uses frequently occurring predicates and identifies the subject and object arguments in the predication, in contrast [21] uses named entity recognition techniques to identify drugs and genes, then identifies the predicates which connect them. This type of 'object-relation-object' inference may also be implied [2]. This method uses 'if then' rules to extract semantic relationships between the medical entities depending on which MeSH headings these entities appear under. For example, if a citation has "Electrocardiography" with the subheading "Methods" and has "Myocardial Infarction" with the subheading "Diagnosis" then "Electrocardiography" diagnoses "Myocardial Infarction".

[28] uses domain-relevant verbs to improve on terminology extraction. The co-occurrence in sentences of selected verbs and candidate terms reinforces their termhood. [29] measures statistical gene name co-occurrence and graphically displays the results for an expert to investigate the dominant patterns.

Question Answering in Biomedicine is surveyed in detail in [31], in particular regarding clinical questions. An example of a system applied to such question is presented in [12], where it is applied in a setting for Evidence-Based Medicine. This system identifies specific 'roles' within the document sentences and the questions, determining the answers is then a matter of comparing the roles in each. To this aim, natural language questions are translated into the PICO format [22]. which is essentially a template of the roles contained in the question. Besides, the identification of roles requires hand written rules which are time consuming to produce and domain specific.

7. CONCLUSIONS

We have described an approach towards automatic extraction of relevant relations in the domain of molecular biology, based on a full parsing of a domain corpus. We have evaluated the performance of the system over a small set of relation of particular interest for the domain expert. The results are extremely encouraging and prompt us to continue in this very promising line of research.

8. REFERENCES

[1] Sophia Ananiadou and Jun'ichi Tsujii, editors. *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, 2003.

[2] J.J. Cimino and G.O. Barnet. Automatic Knowledge Acquisition from Medline. *Methods of Information in Medicine*, 32(2):120–130, 1993.

[3] Michael Collins. *Head-Statistical Models for Natural Language Processing*. PhD thesis, University of Pennsylvania, Philadelphia, USA, 1999.

[4] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*, 1999.

[5] James Dowdall, Fabio Rinaldi, Andreas Persidis, Kaarel Kaljurand, Gerold Schneider, and Michael Hess. Terminology expansion and relation identification between genes and pathways. In *Workshop on Terminology, Ontology and Knowledge Representation, 22-23 January*. Universite Jean Moulin (Lyon 3), 2004.

[6] Steve Finch and Andrei Mikheev. A Workbench for Finding Structure in Texts. In *Proceedings of Applied Natural Language Processing*, Washington, DC, April 1997.

[7] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17(1):S74–S82, 2001.

[8] R. Gaizauskas, G. Demetriou, P. J. Artymiuk, and Willett P. Protein Structures and Information Extraction from Biological Texts: The PASTA System. *Bioinformatics*, 19:135–143, 2003.

[9] Donald Hindle and Mats Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19:103–120, 1993.

[10] Stephen Johnson, editor. *Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical domain*, 2002.

[11] J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182, 2003.

[12] Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. Answering clinical questions with role identification. In Sophia Ananiadou and Jun'ichi Tsujii, editors, *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 73–80, 2003.

[13] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17(2):155–161, 2001.

[14] J. Pustejovsky, J. Castaño, J. Zhang, B. Cochran, and M. Kotecki. Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pacific Symposium on Biocomputing*, 2002.

[15] Fabio Rinaldi, James Dowdall, Michael Hess, Jeremy Ellman, Gian Piero Zarri, Andreas Persidis, Luc Bernard, and Haralampos Karanikas. Multilayer Annotations in PARMENIDES. In *The K-CAP2003 workshop on "Knowledge Markup and Semantic Annotation"*, October 2003.

[16] Fabio Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, and Magnus Karlsson. The Role of Technical Terminology in Question Answering. In *Proceedings of TIA-2003, Terminologie et Intelligence Artificielle*, pages 156–165, Strasbourg, April 2003.

[17] Fabio Rinaldi, James Dowdall, Michael Hess, Kaarel Kaljurand, Mare Koit, Kadri Vider, and Neeme Kahusk. Terminology as Knowledge in Answer Extraction. In *Proceedings of the 6th International Conference on Terminology and Knowledge Engineering (TKE02)*, pages 107–113, Nancy, 28–30 August 2002.

[18] Fabio Rinaldi, James Dowdall, Kaarel Kaljurand, Michael Hess, and Diego Mollá. Exploiting paraphrases in a question answering system. In *The ACL-2003 workshop on Paraphrasing (IWP2003), July 2003, Sapporo, Japan.*, 2003.

[19] Fabio Rinaldi, James Dowdall, Gerold Schneider, and Andreas Persidis. Answering Questions in the Genomics Domain. In *The ACL 2004 workshop on Question Answering in Restricted Domains*, 2004. Accepted for publication.

[20] Fabio Rinaldi, Michael Hess, James Dowdall, Diego Mollá, and Rolf Schwitter. Question answering in terminology-rich technical domains. In Mark Maybury, editor, *New Directions in Question Answering*. MIT/AAAI Press, 2004.

[21] T.C. Rindflesch, L. Tanabe, J. N. Weinstein, and L. Hunter. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing*, pages 514–25, 2000.

[22] D. L. Sackett, S. E. Straus, W. S. Richardson, W. Rosenberg, and R. B. Haynes. *Evidence Based Medicine: How to Practice and Teach EBM*. Churchill Livingstone, 2000.

[23] Eric SanJuan, James Dowdall, Fidelia Ibekwe-SanJuan, and Fabio Rinaldi. A symbolic approach to automatic multiword term structuring. *Computer Speech and Language*, 2004. submitted.

[24] Gerold Schneider. Extracting and Using Trace-Free Functional Dependencies from the Penn Treebank to Reduce Parsing Complexity. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Sweden, November 14-15 2003.

[25] Gerold Schneider, James Dowdall, and Fabio Rinaldi. A robust and hybrid deep-linguistic theory applied to large scale parsing. In *COLING-2004 workshop on Robust Methods in Analysis of Natural language Data, August 2004, Geneva, Switzerland.*, 2004. Accepted for publication.

[26] Gerold Schneider, Fabio Rinaldi, and James Dowdall. Fast, deep-linguistic statistical minimalist dependency parsing. In *COLING-2004 workshop on Recent Advances in Dependency Grammars, August 2004, Geneva, Switzerland.*, 2004. Accepted for publication.

[27] T. Sekimizu, H. Park, and J Tsujii. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Informatics, Universal Academy Press.*, 1998.

[28] Irena Spasić, Goran Nenadić, and Sophia Ananiadou. Using domain-specific verbs for term classification. In Sophia Ananiadou and Jun'ichi Tsujii, editors, *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 17–24, 2003.

[29] B.J. Stapley and G. Benoit. Bibliometrics: information retrieval and visualization from co-occurrence of gene names in MedLine abstracts. In *Proceedings of the Pacific Symposium on Biocomputing (Oahu, Hawaii)*, pages 529–540, 2000.

[30] Martin Volk. Combining unsupervised and supervised methods for PP-attachment disambiguation. In *Proceedings of COLING 2002, Taipeh*, 2002.

[31] Pierre Zweigenbaum. Question answering in biomedicine. In *Proc. of EACL 03 Workshop: Natural Language Processing for Question Answering*, Budapest, 2003.